ID4Fusion@CASE 2025: A Multimodal Approach to Hate Speech Detection in Text-Embedded Memes Using ensemble Transformer based approach

Tabassum Basher Rashfi, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
rashfi2004@gmail.com, {u1904077, u1804128}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

Identification of hate speech in images with text is a complicated task in the scope of online content moderation, especially when such talk penetrates into the spheres of humor and critical societal topics. This paper deals with Subtask A of the Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025. This task is binary classification over whether or not hate speech exists in image contents, and it advances as Hate versus No Hate. To meet this goal, we present a new multimodal architecture that blends the textual and visual features to reach effective classification. In the textual aspect, we have fine-tuned two state-of-the-art transformer models, which are RoBERTa and HateBERT, to extract linguistic clues of hate speech. The image encoder contains both the EfficientNet-B7 and a Vision Transformer (ViT) model, which were found to work well in retrieving image-related details. The predictions made by each modality are then merged through an ensemble mechanism, with the last estimate being a weighted average of the text- and image-based scores. The resulting model produces a desirable F1-score metric of 0.7868, which is ranked 10 among the total number of systems, thus becoming a clear indicator of the success of multimodal combination in addressing the complex issue of self-identifying the hate speech in text-embedded images.

1 Introduction

The emergence of online platforms and social media has changed the channels of communication and sharing of ideas basically. At the same time, this unprecedented liberty of speech has triggered a worrying rise in online hate speech—a message through which a person or group of people are verbalized and violated because of their identity (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Hate speech, especially that carried in the form of images embedded in text (memes), presents a significant challenge for content moderation and online discourse (Gomez et al., 2020). This combination of the textual and the visual mode of presentation makes detection extremely difficult because in many cases when both textual and visual contents are taken together, they can greatly alter their meaning. With the growing complexity of the phenomenon, there has been a trend of automated hate speech detection in research, which has identified both possible applications and limitations in this area, particularly in Natural Language Processing (NLP)-based methods (Parihar et al., 2021).

It is even more difficult to detect hate speech when humor, satire, or coded language are used in memes to mask hateful intentions. The combination of cultural background and rapidly adapting trends online makes it even more difficult and necessitates the usage of multimodal

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 139–145, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-017

explanations that encompass both explicit and implicit indications. The Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE25, in particular, subtask A, which aims to detect the existence of hate speech in images embedded in texts, is a relevant site of discussing these issues (Thapa et al., 2023). Recent work shows that large language models (LLMs) are reshaping computational social science and discourse analysis, while posing key methodological and ethical challenges (Thapa et al., 2025b).

This work presents our approach to Subtask A, where we combine state-of-the-art models from both Natural Language Processing (NLP) and Computer Vision (CV) to create a multimodal system for hate speech detection. The framework has three elements:

1. The textual features are extracted using RoBERTa and HateBERT transformers and classified.

2. EfficientNet and Vision Transformer (ViT) models are used to study the visual content of an image, allowing the identification of visually harmful or offensive content.

3. A scheme of ensemble learning is used to combine the predictions made by any individual modality and thus makes use of complementary information across domains and enhances accurate overall predictions.

The following GitHub repository contains the complete implementation details: https://github.com/RashfiTabassum/Multimodal-Hate-Speech-Detection/tree/main.

2 Related Works

Research on the detection of hate speech in multimodal settings has gone down several approaches with their own limitations. (Pamungkas et al., 2020) achieved 75-80% precision in misogyny detection through the use of machine learning but without visual data. (Derbentsev et al., 2022) concentrated only on text-based methods, whereas (Fortuna and Nunes, 2018) acknowledged that there are few powerful multimodal techniques. (Rawat et al., 2024) explored recent trends, but their techniques struggled with diverse linguistic and visual contexts, reducing generalization. (Kiela et al., 2021) reached an F1 score of 0.80 with the Hateful Memes dataset, which dealt with problems of contextual heterogeneity and uneven distributions. (Cuervo and Parde, 2022) Cuervo and Parde used CLIP to do standardization but had a problem of OCR noise and low flexibility. (Jahan and Oussalah, 2023) restricted their systematic review to NLP-only detection. Meanwhile, (Aluru et al., 2025) introduced a deep-learning framework, yet dependence on the unbalanced information and non-described fusion methods limited its universality.

The CASE shared works have contributed a lot in this field. CASE 2023 (Thapa et al., 2023) was focused on the Russia-Ukraine crisis through the CrisisHateMM dataset, with new subtasks related to hate speech identification and target identifications with multimodal fusion. The scope of CASE 2024 (Thapa et al., 2024) was extended to radicalism, adopting transformer-based NLP and vision models like CLIP and ViT with fusion mechanisms to take into account context, bias, and covert hate such as humor and sarcasm. These two shared tasks provided the foundation for our study.

3 Task and Dataset Description

The Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025) has three differ-

ent datasets in Subtask A: Detection of Hate Speech. It contains 3,662 images, including 1,732 hate images and 1,930 no-hate images. The validation entails 506 images, including 248 labelled as Hate and 258 labelled as No Hate. The test set consists of 507 images, and 249 of them are labeled as Hate, whereas 258 belong to No Hate.

Table 1: Distribution of data for Hate and No Hate categories.

Sets	Hate	No Hate	Total
Train	1985	2065	3662
Val	248	258	506
Test	249	258	507

The dataset is based on the Memeclip corpus (Shah et al., 2024) and the CrisisHateMM (Bhandari et al., 2023) dataset, whose annotation schema was modified for this task. These are the core of the CASE 2025 dataset curation.

4 Methodology

The task objective is to determine the occurrences of hate speech in images embedded with text; thus, a multimodal deep learning methodology to be able to utilize the interaction between the visual and linguistic domains is required. To do this, our method uses a multimodal deep learning architecture that combines CNN-based models for images and pretrained transformer models for text, then employs a fusion strategy that capitalizes on the advantages of both modalities.

4.1 Preprocessing

The preprocessing of the textual data is done by removing the URLs, mentions, non-ASCII characters, digits, and excessive white spaces; all tokens will be automatically transformed to lowercase. Comments that are empty are substituted with an already defined placeholder. Pictures will be resized to 224 x 224 pixels and augmented (additional attempts) by rotations, flipping horizontally, color jittering, and cropping of random parts of images. The images are center-cropped, and the statistical parameters of ImageNet are used to normalize them in a consistent way during validation and testing.

4.2 Text-Based Modeling

For the text modality, we fine-tune two pretrained transformer models, RoBERTa-base and HateBERT (GroNLP), to classify text as either Hate or No Hate. They then tokenized the input text via their respective RoBERTaTokenizer and HateBERTTokenizer with their total length truncated to a maximum of 256 tokens. The AdamW optimizer was used with the learning rate of 1×10^{-5}, and training was done in seven epochs. To handle class imbalance, the class weights were calculated using the scikit-learn function compute_class_weight. To get the final probability of prediction of the text modality, the results of both fine-tuned models were averaged:

$$\text{TextProb} = 0.5 \times \text{RoBERTa} + 0.5 \times \text{HateBERT}.$$

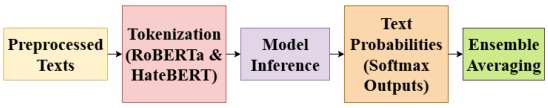

Figure 1: Unimodal Architecture for Text Classification using RoBERTa and HateBERT, followed by Ensemble Averaging.

4.3 Image-Based Modeling

In the case of the image modality, we applied two convolution-based and transformer-based networks: EfficientNet-B7 and Vision Transformer (ViT-B/16). Both of them were pretrained on ImageNet and then fine-tuned on the target dataset, where data augmentation, i.e., horizontal flipping, rotation, color jittering, and random cropping, was used to enable

them to generalize better on unseen data. The training was carried out in 7 epochs using the Adam optimizer at the rate of 1×10^{-5} and 1×10^{-4} of ViT and EfficientNet, respectively. After convergence, the models produced probabilities at the class level; the individual ones were averaged to arrive at the final image prediction:

$$\text{ImageProb} = 0.5 \times \text{EfficientNet} + 0.5 \times \text{ViT}.$$

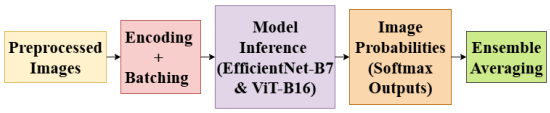

Figure 2: Unimodal Architecture for Image Classification using EfficientNet-B7 and ViT-B16, followed by Ensemble Averaging.

4.4 Multimodal Fusion

We use a late-fusion architecture considering complementary textual and visual data. Textual modalities provide, on average, stronger cues to hate speech as analyzed through the validation procedure, which is empirical. In order to balance the two modalities, we used experiments that changed weights of text-image elements, viz. (0.5, 0.5), (0.7, 0.3), (0.8, 0.2) and (0.9, 0.1). These weight configurations were systematically tested on the validation set in terms of Accuracy, Macro F1, ROC-AUC and class-wise F1 scores. The weighting scheme with 0.7 and 0.3 respectively to textual and visual modality returned the highest Macro F1 and was thus used as the final weighting. The resulting fusion is given as:

$$\text{FinalProb} = 0.7 \times \text{TextProb} + 0.3 \times \text{ImageProb}.$$

Predictions are made by applying a 0.5 threshold on the final probability, classifying the image as Hate or No Hate.

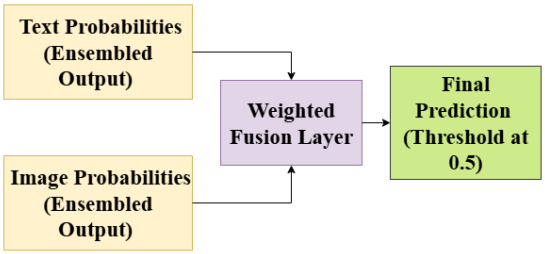

Figure 3: Fusion Layer for Multimodal Prediction using Text and Image Probabilities with Final Thresholding.

5 Experiments and Results

The comparative outputs of various models in terms of macro-averaged Precision (Pr), Recall (Re), and F1-score (F1) have been provided in Table 2. RoBERTa became the best among the text-based models with a macro F1-score of 0.7505, exceeding the results of HateBERT 0.7494. Additional improvement of precision to 0.7990 was made by the ensemble model (RoBERTa + HateBERT), which shows the high ability to combine both models to achieve greater performance. In the image-based models, ViT performed better compared to EfficientNet-B7 with a Macro F1-Score of 0.6351, which is higher than 0.5757 obtained by EfficientNet-B7. The combination of EfficientNet-B7 & ViT had a Macro F1-Score of 0.6311, which shows that two models are more advantageous. The Multimodal Fusion Model, which unites RoBERTa, HateBERT, EfficientNet-B7, and ViT using weights of 70 percent text and 30 percent image, was able to surpass all past models by a big margin. Among all the classification models, the Fusion Model generated the max value of Macro F1-Score (0.7868), Precision (0.7870), and Recall (0.7868).

Table 2: Performance Comparison of Unimodal and Multimodal Models on the Test Dataset

Classifier	Precision	Recall	Macro F1-Score
Unimodal (Text)			
RoBERTa	0.7709	0.7028	0.7505
HateBERT	0.7460	0.7430	0.7494
Ensemble (RoBERTa + HateBERT)	0.7990	0.6546	0.7466
Unimodal (Image)			
EfficientNet-B7	0.5691	0.5622	0.5757
ViT (Vision Transformer)	0.6212	0.6586	0.6351
Ensemble (EfficientNet-B7 + ViT)	0.6220	0.6345	0.6311
Multimodal (Late Fusion)			
Fusion of RoBERTa, HateBERT, EfficientNet-B7, and ViT (70% Text, 30% Image)	0.7870	0.7868	0.7868

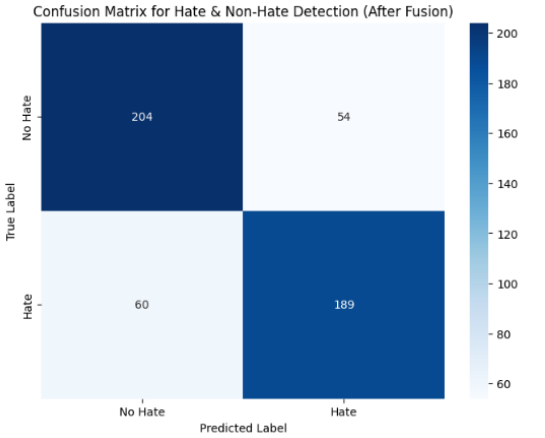

Figure 4: Confusion Matrix for Hate and Non-Hate Detection after Multimodal Fusion

6 Error Analysis

Figure 4, a confusion matrix, indicates some essential misclassification patterns and gives many insights concerning the behavior of the model and its limitations. The multimodal fusion model shows strong results (204 total true negatives and 189 true positives), but there is a tendency to misclassify "No Hate" content as "Hate" (54 false positives) and "Hate" content as "No Hate" (60 false negatives). Such mistakes indicate that the model fails to differentiate between subtle differences in hate speech and other non-hate content. The fact that the false positive rate is relatively high suggests that there might be an over-prediction of hate speech by the model, including instances when surface-level indicators of text and images, such as aggressive words or other visual markings that appear harmful but are not, lead to incorrect predictions. Misclassifications could also be connected to the fact that the model has trouble recognizing humor, satire, or irony, particularly in memes. False negatives emphasize the difficulty of identifying subtle hate speech, including microaggressions and coded speech, which require more context.

7 Conclusions

In our study, we designed a multimodal fusion approach to identify hate speech in images with text, reaching an F1 score of 0.7868, ranking among the top 10 of all systems in the Multimodal Hate Detection Subtask A Shared Task at CASE2025. Fine-tuning the state-of-the-art models such as RoBERTa, HateBERT, EfficientNet-B7, and ViT helped the model take both text and image features into consideration when the model was classifying them to increase the accuracy. Although these are impressive performances, the model suffered with both false positives and false negatives, mainly because it relied on superficial clues and was unable to pick up more subtle manifestations of hate speech, like microaggressions. The following suggestions are intended to improve the context of the subject and the training data and include adding explainability mechanisms to the model to improve precision and minimize false classifications. The paper suggests the possibilities of using multimodality in the identification of hate speech and sets the framework for future developments.

Limitations

We have a number of limitations in our approach. First, the model has trouble identifying subtle and implicit expressions of hate speech, such as microaggression and coded language, because it uses only superficial cues in both text and images. These cues are good against hate speech done on the surface but fail at calling out nuanced forms that need a more in-depth contextualization. Second, the data set is well balanced, but little diversity is provided in hate speech examples that might restrict the model application to generalizing real-world data. Finally, the visual representations used to extract features of images, such as EfficientNet-B7 and ViT, may overlook evolving or symbolic visual symbols in memes and reduce the performance of the model to capture dynamic hate speech.

To address these issues, future improvements could include -

- Increasing the capacity of the model to pick up contextual and implicit cues, perhaps using attention control or context-sensitive fusion.
- Increasing the training data to also have more varied and nuanced data points of hate speech
- The application of explainability tools such as LIME or SHAP might assist in recognizing and correcting these mistakes, thereby resulting in increased precise classification and a lower number of false positives and negatives.

References

Sai Saketh Aluru et al. 2025. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, 15(1):1–15.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Maria Cuervo and Yvonne Parde. 2022. Clip for all: A resource to standardize clip-based research using publicly available data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 769–778. Association for Computational Linguistics. Title adjusted; original title "Exploring Contrastive Learning for Multimodal Detection of Misogynistic Memes" not found in SemEval-2022. Verify with authors if different.

Mykhailo Derbentsev et al. 2022. Deep learning for hate speech detection: A comparative study. *arXiv preprint arXiv:2202.09517*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1470–1478.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

M. S. Jahan and M. Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Douwe Kiela et al. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Endang Wahyu Pamungkas et al. 2020. Misogyny detection in twitter: a multilingual and cross-

domain study. *Information Processing & Management*, 57(6):102360.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Rachana Rawat et al. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. pages 17320–17332.

Surendrabikram Thapa, Farhan Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 151–159, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025b. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.